OPINION ARTICLE

Related to other papers in this special issue	5 (p47); 9 (p87); 6 (p56); 21 (p208); 28 (p276)
Addressing FAIR principles	F, A, I, R

The Need of Industry to Go FAIR

Herman van Vlijmen[1], Albert Mons[†2], Arne Waalkens[3], Wouter Franke[4], Arie Baak[5], Gerbrand Ruiter[6], Christine Kirkpatrick[7], Luiz Olavo Bonino da Silva Santos[8], Bert Meerman[9], Renger Jellema[10], Derk Arts[11], Martijn Kersloot[11], Sebastiaan Knijnenburg[11], Scott Lusher[1], Rudi Verbeeck[1] & Jean-Marc Neefs[1]

[1]Janssen Pharmaceuticals, Antwerpseweg 15, 2340 Beerse, Belgium

[2]Phortos Consultants, The Netherlands

[3]Accenture, Gustav Mahlerplein 90, 1082 MA Amsterdam, The Netherlands

[4]Zorg Instituut Nederland, Willem Dudokhof 1, 1112 ZA Diemen, The Netherlands

[5]Euretos, Yalelaan 1, 3584 CL Utrecht, The Netherlands

[6]Mobiquity, Tommaso Albinonistraat 9, 1083 HM Amsterdam, The Netherlands

[7]UCSD and National Data Service, 10100 Hopkins Dr, La Jolla, CA 92093, USA

[8]GO FAIR International Support & Coordination Office (GFISCO), Leiden, The Netherlands

[9]GO FAIR Foundation, Rijnsburgerweg 10, 2333 AA Leiden, The Netherlands

[10]DSM Biotechnology Center, Alexander Fleminglaan 1, 2613 AX Delft, The Netherlands

[11]Castor, Paasheuvelweg 25, Vleugel 5D, 1105 BP Amsterdam, The Netherlands

Keywords: FAIR application

Citation: H. van Vlijmen, A. Mons, A. Waalkens, W. Franke, A. Baak, G. Ruiter, ... & J.-M. Neefs. The need of Industry to go FAIR. Data Intelligence 2(2020), 276–284. doi: 10.1162/dint_a_00050

ABSTRACT

The industry sector is a very large producer and consumer of data, and many companies traditionally focused on production or manufacturing are now relying on the analysis of large amounts of data to develop new products and services. As many of the data sources needed are distributed and outside the company, FAIR data will have a major impact, both by reducing the existing internal data silos and by enabling the

† Corresponding author: Albert Mons (E-mail: albert.mons@go-fair.org, ORCID: 0000-0001-8038-7572).

efficient integration with external (public and commercial) data. Many companies are still in the early phases of internal data "FAIRification", providing opportunities for SMEs and academics to apply and develop their expertise on FAIR data in collaborations and public-private partnerships. For a global Internet of FAIR Data & Services to thrive, also involving industry, professional tools and services are essential. FAIR metrics and certifications on individuals, data, organizations, and software, must ensure that data producers and consumers have independent quality metrics on their data. In this opinion article we reflect on some industry specific challenges of FAIR implementation to be dealt with when choices are made regarding "Industry GOing FAIR".

1. INTRODUCTION AND CONTEXT

Although the FAIR movement originated in the academic domain [1], industry has responded quickly realizing that they have similar data challenges, and in fact an early IMI project, Open PHACTS [2] was amongst the pioneers leading to the formulation of the FAIR guiding principles. Now, partners from many other industries have embarked on FAIR projects [Janssen, Novartis, Roche, Bayer, Deloitte, Castor, Euretos, Accenture, KPMG, Phortos Consultants, Taxonic, etc.] and IMI has two FAIR dedicated projects running [3]. These days traditional borders between industries are no longer a boundary to growth in the increasingly digital and data driven society. New partnerships arise because no single organization can deliver all the value required for servicing the "smart planet". Businesses must make cross disciplinary connections, breakdown data silos and strive for synergy, using cutting-edge methodologies and tools to identify, develop and deliver new value. An emerging challenge is that data in companies that are merging or starting to collaborate are as difficult to find, access, interoperate, and thus reuse, across different partners as appeared to be the case in academia.

2. THE VALUE OF FAIR DATA

Research data is one of the most valuable resources we have in the world, as it is the key ingredient to innovation, ultimately leading to societal benefits, like alternative energy options, or treatments of diseases. Every element of data could potentially contain a clue that can lead to an important discovery. However, in industry, much like in academia, research data is rarely leveraged beyond its original intended purpose [2]. This is not only based on deliberate data protection, but also on a lack of findability. That means that making data FAIR in industry, and ensuring interoperability and reusability presents a huge opportunity for industry, but ultimately also for society as a whole.

So what is the real value of FAIR data in this context? FAIR data is in itself an abstract concept and runs the risk of being perceived mainly in an IT centric context that is too far removed from end users in the company who should be its principal beneficiaries. Value can only be determined and experienced in a concrete, user specific context. An important success factor in the proliferation and adoption of FAIR data standards is therefore the availability of use cases that clearly demonstrate its value. An industrial example

from the life sciences is the commercially available Euretos AI Platform[①], a solution that aims to address the added value of machine-actionable data through an integration of currently over 250 public life sciences data and textual sources, interlinking multi omics data to scientific literature, patents, experimental and clinical evidence. This is also a pioneering example, demonstrating that even when the company-internal data are not FAIR, but well structured, and machine readable, they can be made FAIR when exported for external use. Thus, company-internal data can be seamlessly combined at request with public data resources. Using this approach, in a recent publication by [4] a machine learning (ML) model was presented that predicts with 78% accuracy whether a particular drug is efficacious for a particular disease, representing an increase of 12% points on earlier state-of-the-art models [4]. The commercial Euretos AI Platform was used in this study and the granular relation types (i.e., predicates) with integrated provenance of over 265,000 direct relations and over 50 million indirect associations through the graph network contributed greatly to this increased performance, showing the added value of FAIR, machine-actionable data.

Another example of the value of FAIR data is a recent Nature survey by [5] which describes several emerging tools like semantic scholar [6] and Iris.ai [7] and also highlighted the Euretos AI Platform in the area of hypothesis generation by end users, including the identification of candidate diseases that existing drugs might treat and identification of gene-expression changes in a neurological disorder called spinocerebellar ataxia type 3 [6]. Extance summarized the early efforts as "*teaching science to machines*", which touches the core of the FAIR principles [8]. Further integration and FAIR linking of public data with proprietary data will undoubtedly provide great value, also in many other data-intensive sectors than in biomedical life sciences.

3. THE NEED FOR A FAIR PUBLIC PRIVATE PARTNERSHIP (PPP)

Ensuring research data is FAIR is a prerequisite of any data-driven organization, but there is typically a lack of research wide agreements on metadata standards, controlled vocabularies and ontologies. In academia, this chaotic situation has been proliferating for many decades and that has resulted in a highly diverse data landscape. For data intensive industries, the rising need for the combination of internal with external data has prompted a need for professional tooling and support in this area and hence increased the need for collaboration for the sectors in academia that now focus on FAIR data and services. For many companies, their core business is in the generation and use of specific data and not in the proper formatting and stewardship of data, nor in the connection to outside data. Thus, here we clearly separate the industrial sector focusing on supporting academia and other industries in professional data FAIRification and the use of FAIR data for research and innovation purposes, from the other sectors who "consume" data FAIRification services from other companies and use them for their specific research, innovation and development purposes. Identifying relevant data sources for model building and meta-studies is facilitated using common naming conventions and standards, as well as the ability to track provenance of data, experimental conditions and broader annotation, which all are core to FAIR. However, industry cannot embark on such

[①] https://www.euretos.com/.

conventions in isolation, as they will create more standards and yet more silos. Therefore, an increasing number of industrial partners is actively engaging in public private partnerships and requiring professional FAIRification support, which creates a new market by itself.

There are numerous FAIR data challenges shared by both commercial and academic organizations, especially related to common vocabularies and ontologies, but there are also differences to be considered. For example, the drive to automate data generation, supported by robotization, parallelization and miniaturization of experiments, although not unique to industry, is an increasingly important component of corporate R&D strategies, and dictates the need for FAIR strategies to be scaled, professionalized and automated with minimal human intervention. This suggests a critical need to stimulate the emergence of professional organisations (SMEs and larger industrial players) able to provide scalable and professionally supported tooling and services in this space. In addition to industrialized internal data generation, collaboration with Contract Research Organisations and academic partners increases the variety and sources of data to be managed, complicated by the fact that they are being created by multiple organizations, each potentially working to their own metadata standards, formats and processes. The high volume of data also results in decentralized storage, with cloud-based strategies an increasingly cost-effective proposition, but requiring data stewardship solutions to specifically address findability and accessibility across multiple disconnected platforms, as well as additional security and privacy concerns.

Ontological disconnect and the associated data modeling are also a significant part of the non-interoperability challenge. There are too many ontologies and knowledge models to choose from, making the semantic modelling of integrated research data a time-consuming task. For example in the life sciences, BioPortal contains 765 ontologies with a total of 9,238,120 classes [9]. A lot of these classes or ontology concepts overlap and share the same label, but there are slight differences in the definition of concepts, and mappings between several ontologies frequently either do not exist or are not complete/exhaustive. Choosing the correct concept, especially for researchers that have little to no experience with ontology mapping, is challenging and performing ontology mapping for a complete data set is an extensive task, which again is a shared challenge with academia.

A related but different challenge is the semantic context and provenance of the data: how can different knowledge models be correctly mapped [10], how was the data measured, what was the experimental design? Even if the same ontologies are used in different datasets, the experiments may still be measuring slightly different things with slightly different settings or methods, which could cause massive reproducibility problems and misinterpretations. Rich provenance type of metadata is often not available or briefly described as free text. There are some efforts to standardize context descriptions, e.g. in the minimal information standards (such as MIAME for gene expression data) [11] but much more is needed.

4. BENEFITS FOR DATA INTENSIVE INDUSTRY

Pharmaceutical industry is an example of a data intensive industry and for the purpose of this article we briefly zoom in on Pharma related data aspects, although most comments are equally true for other

industries. Drug discovery is difficult because many of the properties required by a drug, especially efficacy and toxicity in man, cannot be predicted well. Because of this lack of predictability it is often said that developing a new drug is more complex than developing a new airplane. Any improvement in predictive models can have a major impact on drug discovery, and therefore access to all available data is critical.

In many areas proprietary corporate data is much more detailed than public data, for instance: pharmacology assay data, data on absorption, distribution, metabolism, and excretion (ADME), toxicity data, and clinical data of specific drug candidate molecules. Public data, on the other hand, tend to be much broader, for instance in genetics and various omics databases. For the true realisation of personalised medicine, large amounts of data are needed, especially in relation to outcome predictions (Value Based Health Care). Ideally this will include integration of multiple proprietary with public data sources in a way that preserves the ownership of the data while allowing specific analyses to be done [12].

In reality however, many internal databases in pharma are in silos and they are often much less FAIR than many public databases [2]. One of the reasons for this is that until recently most generated data was used inside specific discovery/development projects, and the need or interest for sharing and reuse was low. Data reuse has become much more important for developing predictive models, learning from past experience, and therefore many pharmaceutical companies are actively improving the reusability of data. This situation inspired 7 pharma companies to initiate the IMI call for FAIRification of data, resulting in the public-private FAIRplus consortium [3], which aims to increase the discovery, accessibility and reusability of data from selected IMI projects and internal data from pharmaceutical industry partners". It will also organise training for data scientists in academia, SMEs and pharmaceutical companies to enable wider adoption of best practices in life science data management. We expect that this example will soon be followed in many other industrial sectors.

Time is of the essence in all research and is particularly important in the highly competitive industries. In Machine Learning and increasingly in early Artificial Intelligence approaches the access to large quantities of quality data, and this includes all aspects of FAIR, is often seen as more important than the algorithms used, especially with the emergence of tools that automatically select the best possible algorithm for the task given the available data. During recent years, the quantity of available data has increased not just by proper vocabularies and metadata, but also by access improvement (technically (API) and by clarifying access restrictions).

It is important for industry that FAIR (Accessible) data is not the same as Open data. Even within a company not all data can be used by all scientists, and finding this out often means finding the right person to talk to, which can be very time consuming. So, besides making it technically possible to access certain data, it is important to have clear access criteria, thereby providing scientists and machines with all the data they have a right to use, in a timely manner.

5. BENEFITS TO FAIR DATA SERVICE PROVIDERS

Increasingly the world is starting to appreciate the value of FAIR data and being a provider in this space while these trends are emerging is a huge competitive advantage. On the one hand this is because professional services can be offered that leverage FAIR data and on the other hand because it allows for bringing in accounts that have a future need for FAIR data and want to work with a provider that understands that need and can deliver on it.

In addition, the trend is that FAIR data is highly needed in innovation which in turn increases the need for new products supporting the analytics possible with machine readable data. While there is no protocol or standard to make data FAIR (yet), although several efforts are underway [13], adapting the FAIR principles early on allows industry (in combination with academic institutes) to set a standard or ”best practice”.

Another interesting benefit in a world where high quality data experts are scarce is employee engagement by explaining to co-workers that working on standardizing the world’s research data is a huge benefit for society and also an interesting and ever evolving field. The current generation of workers cares deeply about making an impact, and being a FAIR service provider means one can easily explain how the company is making a positive impact on the world.

Lastly, everyone recognizes the power and value of data, being able to profile the company as leading the charge towards an interoperable world has a big positive impact when speaking to governmental organizations, funding agencies and investors.

6. CURRENT LAY OF THE LAND FAIR TOOLING AND SERVICES

Over the last several years many FAIR tools have been developed, mainly as prototypes or reference implementations, often, tongue-in-cheek referred to as ”Professorware” [14].

As any technological advance, in order for FAIR to reach mainstream, industry needs to have available a number of professional products and services to support the creation and use of FAIR data. Currently, the process to FAIRify data includes a number of manual steps that can and should be better supported and even fully automated. The main goal is to relieve as much as possible, from one side, the burden to provide FAIR data and, from the other side, to consume them. With this support, and consequent reduction in the required effort, more resources will be available to provide more FAIR data, from the producer’s perspective, and to find, access, interoperate and ultimately reuse more data, from the consumer’s perspective.

An interesting development is the increasing request from ”FAIR customers” for FAIR as a Service (FaaS). This would provide for a multi-tenant cloud-based offering allowing for a seamless interaction between the FAIR tools supporting the whole cycle from planning (for example the Data Stewardship Wizard [15] publication (i.e., a FAIR Data Point [16]) and Evaluation (for instance the growing number of FAIR maturity Evaluators [14]) CKAN being one example for validating data compliance [17]).

7. FAIR DATA AND CERTIFICATION

One way to assure the required quality of the FAIR data and services for use in industry is certification. The certification can be on several levels. The development of the FAIR Data Stewardship skill certification for individuals is an important first one. The second level is the certification of data sets. By measuring and certifying the ″FAIRness″ of a data set, the ″principle usefulness″ of the data set can be assessed. The third and fourth level of certification are for organisation and software. Especially the latter can be useful in developing interoperable architectures based on IFDS (Internet of FAIR Data and Services).

Since the ″market″ (both academia and industry) increasingly asks for a quality seal for FAIR data, and there is a sense that "trust in data received" is currently a bottleneck in user acceptance, the GO FAIR Foundation (GFF) [18], supported by the Dutch Ministry of Economic Affairs, is developing a first generation FAIR Certification Programme that aims to work across different industries and different countries. It is GFF's goal to increase the trust-factor in projects where data is shared between several (scientific and/or commercial) parties. With a trusted, neutral party and a coherent certification program, the implementation time of data exchange projects will be much shorter.

Preparing and implementing a solid certification scheme for FAIR Data will require an approach whereby business, legal and technical expertise needs to be combined to create a coherent program. This FAIR certification program needs to be designed thoroughly to secure its future success in both the academic world and industry. Furthermore, the program has to make sure that it is in line with different governmental and industry policies on data sharing, including of course GDPR.

In order to be successful, and a trusted party for industry, GFF will need to develop relationships with organisations from different industries (for instance finances, logistics, agriculture, medical, chemistry, etc.), and needs to verify which potential partners are interested in implementing the FAIR principles and services in a particular area. Eventually the GFF will need to work with or become itself a player and work with licensed certification bodies giving it the certificates for individuals, data, tools and organizations.

8. THE PUBLIC PRIVATE PARTNERSHIP AS FAIR TRUSTED PARTY

Next to the same improvements in reliability, reproducibility and translational power that is needed for science and innovation in general, data intensive companies need an even stronger ″robustness″ of all elements used in their pipelines, especially when it comes to heavily regulated processes such as drug- or instrument development. Thus, proper licensing, professional versioning, documentation and guaranteed up-time become even more critical than in more exploratory sciences.

For the private sector to move forward on FAIR a couple of prerequisites need to be addressed. At first a trusted party is required to bring together public and private parties. Both are needed to develop successful implementations of IFDS. The public sector can organize incentives and help with regulatory requirements. Private organizations need to provide capacity and knowledge.

Therefore, high quality services, data, and a scaling capacity is essential for industry to optimally profit from the FAIR ecosystem. This "environment" needs to be built with content specialists from academia and industry as well as with professional engineering and service provider industries. However, there will be silos again, unless the services function in a scalable IFDS that can fluidly cross the public-private boundary and can operate on closed as well as open data, as long as both are FAIR.

One of the challenges for the partnership is to address creating enough service capacity in the market for all FAIR initiatives and to maintain quality of the FAIR capacity. Therefore, industry should contribute centrally and intensively to the choice of the common protocols and standards (such as a commonly agreed FAIR Digital Object IP and connector service) that would enable a scalable IFDS facilitating the public-private dovetailing of data and services under well-defined conditions. Ideally all tooling needed for a seamless FAIR process should be provided as a cloud based multi-tenant "FAIR as a Service" (FaaS).

9. THE FAIR SERVICE PROVIDER CONSORTIUM

Looking forward, the FAIR market prospects indicate that there is a need for professional FAIR services. The GO FAIR Foundation and Phortos Consultants have taken the initiative to approach service provider companies in the data realm to form the FAIR Service Provider Consortium (FSPC). To date 10+ companies have joined and agreed to build capacity by training FAIR Data Stewards and Ontologists. Several partners are looking into establishing a FAIR Center of Competence.

A development worth mentioning is that the FSPC has committed to adhere to the GO FAIR Rules of Engagement and implement according to the best practices as they are formulated by the GO FAIR Implementation Networks.

Several FSPC partners have expressed interest in (co) developing professional tooling (as mentioned in the preceding paragraph) to be provided in a FAIR as a Service cloud environment. The services the FSPC provides range from FAIR Awareness events, to FAIRification of data, semantic and ontology modeling, building FAIR compliant tooling, FAIR Data Stewardship training and assist companies in the process of GOing FAIR.

This growing consortium (and hopefully similar ones in the future) will be crucial to systematically build professional support capacity, but also the awareness, emphasize and guide the professionalisation of tools and services and a growing industrial adoption of FAIR principles and their implementation in the industrial sectors.

AUTHOR CONTRIBUTIONS

A. Mons (albert.mons@phortosconsultants.com) designed the outline of the article and wrote a first outline which was reviewed and augmented by H. van Vlijmen (hvvlijme@its.jnj.com) after which all authors wrote sections of the article based upon there specific background and experience. All authors contributed to the

writing and provided critical feedback to help shape the manuscript. In addition, all authors edited and reviewed the final version of the article.

REFERENCES

[1] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak et al. & B. Mons. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.

[2] A.J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E.L. Willighagen ... & B. Mons. Open PHACTS: Semantic interoperability for drug discovery. Drug Discovery Today 17(2012), 1188–1198. doi: 10.1016/j.drudis.2012.05.016

[3] FAIRplus project. Available at: https://fairplus-project.eu/.

[4] W.J. Vlietstra, R. Vos, A.M. Sijbers, E.M. van Mulligen & J.A. Kors. Using predicate and provenance information from a knowledge graph for drug efficacy screening. Journal of Biomedical Semanticsvolume 9(2018), Article No. 23. doi: 10.1186/s13326-018-0189-6.

[5] A. Extance. How AI technology can tame the scientific literature. Nature 561(2018), 273–274. doi: 10.1038/d41586-018-06617-5.

[6] L.J.A. Toonen, M. Overzier, M.M. Evers, L.G. Leon, S.A.J. van der Zeeuw, H. Mei ... & W.M.C. van Roon-Mom. Transcriptional profiling and biomarker identification reveal tissue specific effects of expanded ataxin-3 in a spinocerebellar ataxia type 3 mouse model. Molecular Neurodegener 13(1)(2018), 31. doi: 10.1186/s13024-018-0261-9.

[7] Research discovery with artificial intelligence. Available at: http://iris.ai.

[8] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, ... & E. Schultes. FAIR principles: Interpretations and implementation considerations. Data Intelligence 2(2020), 10–29. doi: 10.1162/dint_r_00024.

[10] Bioportal. Available at: https://bioportal.bioontology.org/.

[11] G. Guizzardi. Ontology, ontologies and the "I" of FAIR. Data Intelligence 2(2020), 181–191. doi: 10.1162/dint_a_00040.

[12] http://fged.org/projects/miame/.

[13] O. Beyan, A. Choudhury, J van Soest, O. Kohlbacher, L. Zimmermann, H. Stenzhorn, Md. R. Karim, M. Dumontier, S. Decker, L.O. Bonino da Silva Santos & A. Dekker. Distributed analytics on sensitive medical data: The Personal Health Train. Data Intelligence 2(2020), 96–107. doi: 10.1162/dint_a_00032

[14] FAIR Data Maturity Model WG. Available at: https://www.rd-alliance.org/groups/fair-data-maturity-model-wg.

[15] B. Mons. Data stewardship for open science: Implementing FAIR principles. Boca Raton: CRC Press, 2018. isbn: 9780815348184.

[16] DS-Wizard. Available at: https://ds-wizard.org/.

[17] FAIR data point. Available at: https://www.research-software.nl/software/fairdatapoint.

[18] CKAN. Available at: https://ckan.org/2018/01/15/introducing-ckanext-validation-data-validation-and-reporting-integrated-in-ckan/.

[19] GO FAIR. Available at: https://gofairfoundation.org/.